多特征知识下的食品安全事件实体抽取研究*

王东波[1,2]　吴　毅[1]　叶文豪[1]　刘睿伦[1]
[1](南京农业大学信息科学技术学院　南京 210095)
[2](南京农业大学领域知识关联研究中心　南京 210095)

摘要: 【目的】从大规模食品安全事件当中抽取食品安全事件实体。【方法】基于已发生的食品安全事件, 结合情报学数据获取、标注和组织的方法, 融合食品安全事件实体的多种分布特征知识, 通过条件随机场模型, 构建食品安全事件语料并从中抽取相应的实体。【局限】在食品安全事件实体抽取过程中所制定的特征模板在领域化迁移上具有一定的局限性。【结果】在已有 1 500 万字经过标注的食品安全事件语料的规模上, 通过统计食品安全事件实体的内部和外部特征, 基于条件随机场机器学习模型, 构建了食品安全实体的抽取模型, 该模型最高的 F 值达到 91.94%。【结论】通过对食品安全事件实体抽取结果的分析, 在食品这一领域化的语料上, 基于条件随机场进行实体抽取是可行的。

关键词: 特征知识　条件随机场模型　实体　食品安全事件
分类号: G350

1　引　言

为了应对备受关注的"双汇瘦肉精"、"老酸奶"、"酒鬼酒塑化剂超标"、"致癌金针菇"、"美素丽儿奶粉"、"硫磺熏制枸杞"、"镉大米"等食品安全事件问题, 2013 年 12 月 23 日至 24 日的中央农村经济工作会议明确提出"尽快建立全国统一的农产品和食品安全信息追溯平台"的具体措施, 而构建食品安全信息追溯平台的基础是要对食品安全事件中的主要实体进行确认, 尤其是涉及到食品安全舆情的处理时, 相关实体的抽取变得愈发重要。针对这一情况, 本文基于构建的食品安全事件语料库, 结合条件随机场机器学习模型, 通过利用食品安全事件实体的多特征知识, 对食品安全事件的实体进行抽取实验。一方面为构建食品安全事件知识库提供了基本的知识锚点, 另一方面也为深入挖掘、分析和总结应对食品安全事件的策略奠定了基础。

有关食品安全事件的研究主要集中在案例、政策和应急处理上, 有代表性的研究主要有: 由复旦大学的研究生吴恒联合 34 名网络志愿者创建"掷出窗外"网站[1], 搜集了关于食品安全事件的相关事件并构建了数据库。该数据库为本文构建的食品安全事件语料库提供了一定数量的文本, 是本文语料库构建的基础。

关于食品安全事件的研究更多是从管理学的角度进行, 比较有代表性的研究有: 张慕洁等[2]基于两个典型案例, 分析了应急管理事件时信息不公开造成的危害, 并探讨了常见的不公开的原因。该研究选取典型案例的方法为本文确定语料文本提供了方法上的借鉴。马颖等[3]构建了食品行业事件风险感知的传染病模型, 并以日本地震衍生的"抢购食盐事件"为例, 对

通讯作者: 王东波, 0000-0002-9894-9550, E-mail: db.wang@njau.edu.cn。

*本文系 2011 湖北省协同创新中心项目"面向应急推演平台的海量突发事件知识库与模型库构建研究"(编号: JD20150101)、国家自然科学基金项目"基于 CSSCI 的句法级汉英平行语料库构建及知识挖掘研究"(项目编号: 71303120)和地震科技星火计划项目"面向地震应急的空间智能决策方法研究"(项目编号: HX15019)的研究成果之一。

模型进行数值分析和检验。该研究为本文进行食品安全事件的名称标注提供了相应的借鉴之处。

上述研究一方面为本文提供了宏观的方法、策略指导，另一方面也为本文确定食品安全事件的实体提供了具体的依据。

实体的抽取方面最新的研究主要是通过机器学习的方法抽取非结构化文本中的实体，比较有代表性的研究如下：基于神经网络的策略，陈宇等[4]尝试利用 Deep Belief Nets 模型对实体及实体之间的关系进行抽取。该研究为本文确定特征量的数量提供了相应的方法指导。利用相应的语义知识对实体进行抽取也是目前较为流行的策略，邵发等[5]从解决一词多义的问题着手，利用歧义消除策略，通过 HowNet 和贝叶斯分类的资源与方法，对实体进行抽取。从消除歧义的角度完成对实体的识别虽然具有一定的科学性，但这种方法在大规模的语料上的整体性能有待验证。针对急剧增加的电子医疗文本，许华等[6]基于分词、词性标注的医疗语料，利用规则的方法，完成对医疗文本中实体的抽取，整体性能达到 80%以上。规则的方法虽然在某一特征的语料上具有一定的适应性，但由于对蕴含在具体语料词汇之间的规则缺乏充分的探究，在一定程度上会导致所制定规则的覆盖度相对较差。这也是本文选取条件随机场模型进行食品安全事件实体抽取的主要原因之一。与食品安全事件相关的信息抽取研究中，目前集中在针对食品投诉文本词汇层级的知识抽取，比较有代表性的研究是魏秀卓[7]围绕食品投诉文本敏感词汇的抽取和高蕊[8]基于本体的食品投诉文本危害信息的提取。相对于实体抽取，词汇级的抽取相对简单，主要体现在词汇的长度较短和内部组成相对简单这两点上。条件随机场作为抽取术语和实体等序列化的机器学习模型具有较广泛的应用，比较有代表性的如下：李丽双等[9]通过简单特征模板完成对汽车术语的抽取；在词汇组合的特征模板基础上，王文龙等[10]完成了对项目申报书中实体的抽取；结合中医词汇的特征知识，刘凯等[11]构建了中医电子病历的实体抽取模型。上述基于条件随机场的术语和实体抽取仅仅利用了实体自身简单的特征知识，未涉及到所抽取对象上下文语境的信息。本文在识别食品安全事件实体的过程中构建了复杂的特征模板，在一定程度上弥补了已有识别方法的不足。

2　食品安全事件实体界定和特征统计

2.1　食品安全事件语料简介及实体界定

在对食品安全事件进行采集、标注和组织的基础上，本文构建了 2005 年–2015 年的食品安全事件语料库。食品安全事件的获取目标主要包括互联网上的食品安全事件和纸质媒介上的食品安全事件。网络上食品安全事件的采集主要通过面向事件主题垂直搜索引擎技术自动采集，采集范围包括新闻门户、论坛和博客，对于采集的异构数据通过相应的数据清洗、转换统计保存到数据库中，而纸质的事件案例则通过人工录入、校对的方式完成对事件的采集。具体食品安全事件文本抓取的程序截图如图 1 所示。

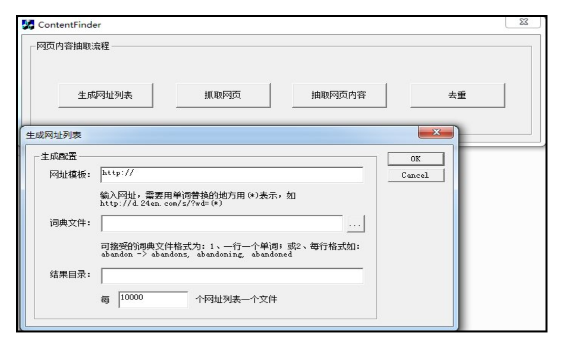


图 1　食品安全事件抓取软件截图

食品安全事件的标注主要是完成对食品安全事件的分词、词性标注，针对词汇长度比较大的食品安全名称则标注大一级的词性，相对通用的语料，食品安全事件中的词汇长度较长，在进行分词的时候将这类词汇视为一个词汇来处理并进行词性标注；食品安全的组织主要是对食品安全事件进行类别标注，具体类别标注则基于《中华人民共和国食品安全法》进行。经过上述处理，所构建的食品安全事件语料库达到 1 500 万字级和 687 万词级，由 2 800 个食品安全事件组成。

本文的实体主要是指食品安全事件中涉及的食品名称与导致食品安全事件发生的具体因素，比如具体的食品名称有“奶粉、酱油、大米、牛奶”等，而具体因素则为“添加剂、甲醛、过氧化苯甲酰、反式脂肪酸”等。本文的主要任务是构建机器学习模型，自动将食品名称与导致食品安全事件发生的具体因素抽出来。条件随机场模型训练和测试所使用的语料样例如下所示。

企业/n 或/c 个人/n 的/u "/w 违法/vn 行为/n "/w 中/f , /w 包括/v "/w 生产/v 假冒/vn 注册/vn 商标/n 的/u 瓶装/b 水/n "/w "/w 在/p 生产/vn 加工/vn 饺子皮/n 、/wn 云吞皮/n 过程/n 中/f 添加/v 有毒/vi 有害/a 物质/n【硼砂/n】"/w "/w 在/p 生产/vn 加工/vn【牛百叶/nr】、/wn【鱿鱼/n】、/wn【牛】肚/ng 等/v 食品/n 的/u 过程/n 中/f 添加/v 有毒/vi 有害/a 物质/n【过氧化氢/n】和/p 氢【氧化钠/n】"/w "/w 递交/v 虚假/a 材料/n 取得/v 餐饮/n 服务/vn 许可/vn "/w "/w 篡改/v 食品/n 生产/vn 日期/n 并/d 销售/v "/w 等/u 。/wj

2.2　实体内部和外部特征统计

选取 2 800 个食品安全事件, 通过人工对其中的食品名称与导致食品安全事件发生的具体因素进行手工标注。在标注的语料基础上, 统计"食品名称"与"具体因素"这些实体的内部和外部特征。

(1) 内部特征

①词语长度

获取实体的长度一方面有利于掌握所抽取实体对象的难易程度, 另一方面也有利于确定条件随机场标记集的数目。食品安全事件实体长度分布如表 1 所示。

表 1　食品安全事件实体长度分布表

实体长度	数量(个)	实体长度	数量(个)
2	48 036	13	13
3	23 499	9	9
4	6 878	10	7
1	6 594	12	5
5	1 383	14	2
6	394	11	1
7	182	15	1
8	37	20	1

由表 1 可以看出, 实体的长度主要在 1-5 之间, 通过计算得出长度为 1-5 的实体占总数的 99.25%, 长度为 2 和 3 的实体占总数的 82.18%, 长度为 2 的实体占总数的 55.19%, 长度为 3 的实体占总数的 27.00%。通过计算结果不难发现: 长度为 2 的实体数量超过半数, 因此在实体抽取方面, 长度为 2 和 3 的实体是重点抽取的对象, 例如"奶粉"、"牛奶"、"猪肉"、"添加剂"、"地沟油"等。而那些长度大于 8 的大多是含有形容词的名词或是一些复杂的专有名词, 例如: "环己基氨基磺酸钠"。

②具体实体的分布情况

通过统计具体实体的分布情况不仅有助于获得感性的有关实体的具体内容, 而且也有利于统计具体实体的左右特征知识。部分食品安全事件实体的分布如表 2 所示。

表 2 只选取了部分实体数据, 分别是排名前 10 以及实体长度为 4-6 的数量靠前的实体数据(该数据共有 3 193 项, 87 042 个)。因为排名前 10 中大部分为长度为 2 的实体, 故未在表格中再添加该类数据。该项统计的实体总量为 87 042, 其中前 10 项占总数的 13.16%, 前 5 项占总数的 8.23%, 第二项奶粉占 1.91%, 第一项添加剂占 2.58%。

表 2　具体食品安全事件实体的分布情况

实体	数量(个)	实体	数量(个)
添加剂	2 243	大米	899
奶粉	1 661	牛奶	810
地沟油	1 178	药袋	733
酱油	1 078	菌落总数	377
酒	1 006	亚硝酸盐	352
猪肉	943	反式脂肪酸	95
甲醛	904	过氧化苯甲酰	90

(2) 外部特征

在不同食品安全事件的语料中, "食品名称"和"具体因素"的左右边界存在较大的差异, 分别对食品安全事件语料中的"食品名称"和"具体因素"的左右边界进行统计, 该统计结果对于后续构建"食品名称"和"具体因素"抽取模型具有重要价值。

"食品名称"和"具体因素"的边界范围限定在以"。! ? "结尾的子句范围内, "食品名称"和"具体因素"的左边界绝对不会跨越其第一个标记即"食品名称"和"具体因素"的起始标记, 因此考察范围限定在从句子开始到第一个标记结束的范围内, 称为 β。同样, "食品名称"和"具体因素"的右边界特征词绝对不会跨越"食品名称"和"具体因素"的最后一个标记, 因此考察范围限定在从最后一个标记开始到句子结束这样一个范围内, 这个范围记做 α。具体选取"食品名称"和"具体因素"左边界词的计算公式如公式(1)[12]所示。

$$P(w)=\frac{f(W_left_outside)}{f(W_left)} \tag{1}$$

其中, $f(W_left_outside)$ 表示 W 在 β 范围内出现的频次, $f(W_left)$ 表示 W 在 β、"食品名称"、"具体因素"内部出现的频次。通过公式(1), 结合食品安全事件的语料, 给定 P 的经验阈值为 0.8, 即当 $P\geq 0.8$ 时, W 可能成为"食品名称"和"具体因素"的左边界词, 然后结合人工语言学知识的内省, 最终确定 7 个左边界词: "的、用、和、是、食品、超标、中"。

同理, 使用公式(2)[12]用于"食品名称"和"具体因素"右边界词的选取。

$$P(w)=\frac{f(W_right_outside)}{f(W_right)} \tag{2}$$

其中，$f(W_right_outside)$ 表示 W 在 α 范围内出现的频次，$f(W_right)$ 表示 W 在 α、“食品名称”、“具体因素”内部出现的频次，将右边界词 P 的阈值也设定为 0.8，根据语言学知识的内省再结合大于或等于 0.8 的 P 值，最终确定 10 个右边界词：“的、用、品、有、种、和、是、超、中、产”。

3 模型简介和特征确定

3.1 机器学习模型

条件随机场是由 Lafferty 等[13]提出的用于解决序列标注问题较优的一种模型，是在给定一组需要标记的观察序列的条件下，计算整个观察序列状态标记的联合条件概率分布的无向图模型。对于指定的节点输入值，能计算指定节点输出值的条件概率，其训练目标是使得条件概率最大化。最常用的 CRFs 模型是一阶链式结构，即线性链结构，其拓扑结构如图 2 所示。

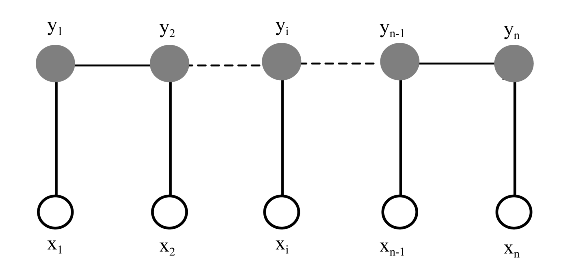


图 2 线性链 CRFs 模型的拓扑结构

设 $x=\{x_1, x_2, \cdots, x_{n-1}, x_n\}$表示被观察的输入数据序列，如本文语料中分词后的词；$y=\{y_1, y_2, \cdots, y_{n-1}, y_n\}$表示有限状态集合，其中每个状态对应于一个标记。在给定输入序列 x 的条件下，对于参数 $\lambda=\{\lambda_1, \lambda_2, \cdots, \lambda_{n-1}, \lambda_n\}$的线性链 CRFs 的状态序列 y 的条件概率如公式(3)和公式(4)所示[13]。

$$p(y\mid x,\lambda)=\frac{1}{Z_x}\exp\left(\sum\nolimits_{i=1}^{n}\sum\nolimits_{j}\lambda_j f_j(y_{i-1},y_i,x,i)\right) \tag{3}$$

$$Z_x=\sum\nolimits_{y}\exp\left(\sum\nolimits_{i=1}^{n}\sum\nolimits_{j}\lambda_j f_j(y_{i-1},y_i,x,i)\right) \tag{4}$$

其中，Z_x 为归一化因子，表示所有可能的状态序列的得分，确保所有可能状态序列的条件概率之和为 1。$f_j(y_{i-1},y_i,x,i)$ 是一个统一形式的特征函数，通常为二值表征函数；λ_j 是通过模型对训练数据进行训练之后获得的相应特征函数的权重。最大熵模型(ME)是以 McCallum 等[14]提出的最大熵原理为基础，即最大熵的原理主要是如果概率分布信息不确定，那么最不会产生偏置的做法，就是均等看待概率分布，不要做任何主观假设，在给定关于训练数据的限制条件下，使模型的熵达到最大的分布，就是所求分布。最大熵模型在人工智能和自然语言处理等领域也得到广泛应用，但由于最大熵模型本身存在标注偏置的问题，错误识别和未识别的情况较多，导致在某些情况下其效果不如 CRF 等模型。

3.2 语料的选择和语料的处理

具体的“食品名称”和“具体因素”的实体在语料中被标注成“【 】”的形式，如：

“【/wky 牛奶/n】/wky, /wd 30/m 余/m 位/q 执法/vn 人员/n 来到/v 西/b 长街/n 农贸市场/n”,核实/v【/wky 反式/b 脂肪酸/n】/wky 的/u/问题/n。/wd

基于对“食品名称”和“具体因素”的特征统计，在条件随机场模型定义基础上，本文在确定用于“食品名称”和“具体因素”的 CRF 标记数的过程中，主要参考公式(5)[14]。

$$L=\frac{1}{N}\sum\nolimits_{i=j}^{k} iN_i \tag{5}$$

其中，L 表示当 $i\leqslant k$ 时“食品名称”和“具体因素”时平均加权后的长度，N_i 表示所选取的语料中长度为 i 的“食品名称”和“具体因素”出现的次数，k 和 j 分别表示语料库中最长与最短“食品名称”和“具体因素”的长度，N 表示语料库中“食品名称”和“具体因素”的总个数。基于公式(5)，结合语料的基本情况以及相应的实验结果，“食品名称”和“具体因素”识别模型构建中确定使用 5 词位的标注集，标注集用 R 来表示，具体为 R={B, C, E, S, A}，B 表示“食品名称”和“具体因素”的初始词，C 为“食品名称”和“具体因素”的中间词，E 为“食品名称”和“具体因素”的结束词，S 为“食品名称”和“具体因素”之外的词汇，A 为一个词或字单独为“食品名称”和“具体因素”的情况，如果“食品名称”和“具体因素”的长度超过 3，就用 C 表示扩展词。

本文通过编写 Java 程序，结合语料中“食品名称”和“具体因素”的“【 】”标记以及根据选取的特征及制定的特征模板，自动对所有语料进行标注，具体标注样例如表 3 所示。

表 3 "食品名称"和"具体因素"训练语料和测试语料标注样例

词语	词性	词长度	是否实体词	是否左边界	是否右边界	标记
有关	p	2	N	N	N	S
反式	b	2	Y	N	N	B
脂肪酸	n	3	Y	N	N	E
问题	n	1	N	N	N	S
,	wd	1	N	N	N	S
浙江省	ns	3	N	N	N	S
金华市	ns	3	N	N	N	S
公安局	n	3	N	N	N	S
江南	ns	2	N	N	N	S
分局	n	2	N	N	N	S
接到	v	2	N	N	N	S
群众	n	2	N	N	N	S
举报	vn	2	N	N	N	S
称	v	1	N	N	N	S

3.3 特征的选取以及特征模板的制定

对于基于条件随机场的机器学习模型中，特征的选择都极其重要。特征选择的好坏将会直接影响到CRFs模型的性能。特征由原子特征和复合特征两部分构成。本文选取的原子特征为词语本身、词性、词长度、是否实体词、是否左边界、是否右边界等6个特征；复合特征是通过对原子特征的组合来表征"食品名称"和"具体因素"实体复杂的语言学特征。6个特征选择的特征窗口大小分别为7,3,5,5,5,5，7个窗口的范围是{–3,–2,–1,0,1,2,3}，5个窗口的范围是{–2,–1,0,1,2}, 3个窗口的范围是{–1,0,1}。在上述特征中，从对"食品名称"和"具体因素"抽取性能提升的角度考虑，词性和词语本身是最重要的特征，其次是左右边界词和实体词，最后是"食品名称"和"具体因素"的长度。

4 实体抽取实验

对抽取模型性能的评价主要采用三个指标来衡量：准确率(Precision)、召回率(Recall)、F值(F-measure)。分别基于上文标注的语料使用条件随机场模型和最大熵模型进行"食品名称"和"具体因素"的抽取。在具体的实验中使用交叉验证的方法测试所构建模型的性能，将2 800个语料文档分别按照9:1的比例分为训练语料和测试语料，测试结果如表4和表5所示，表6展示了两种模型在同样的软硬件条件下训练和测试耗时的对比。

表 4 基于条件随机场模型"食品名称"和"具体因素"抽取性能比较

测试编号	准确率	召回率	F值
1	89.95%	90.17%	90.06%
2	90.46%	91.01%	90.73%
3	91.89%	90.68%	91.28%
4	88.35%	91.88%	90.08%
5	90.37%	91.06%	90.71%
6	91.01%	90.07%	90.54%
7	91.43%	91.74%	91.58%
8	90.48%	91.01%	90.74%
9	92.12%	91.77%	91.94%
10	90.54%	91.65%	91.09%
均值	90.66%	91.10%	90.88%

表 5 基于最大熵模型"食品名称"和"具体因素"抽取性能比较

测试编号	准确率	召回率	F值
1	72.55%	62.50%	67.15%
2	73.72%	61.89%	67.29%
3	81.90%	65.19%	72.60%
4	84.10%	59.97%	70.01%
5	81.67%	62.49%	70.80%
6	86.52%	63.70%	73.38%
7	81.66%	65.74%	72.84%
8	72.71%	67.10%	69.79%
9	74.72%	63.37%	68.58%
10	80.88%	65.40%	72.32%
均值	79.04%	63.74%	70.48%

表 6 条件随机场和最大熵模型训练和测试耗时比较

编号	条件随机场模型		最大熵模型	
	训练耗时(秒)	测试耗时(毫秒)	训练耗时(秒)	测试耗时(毫秒)
1	43 837.09	810	78.01	4
2	41 660.11	1 045	67.01	5
3	43 267.72	980	89.06	78
4	42 078.04	124	67.35	9
5	41 863.00	450	56.43	45
6	43 287.12	160	67.50	7
7	45 677.87	678	57.49	67
8	48 814.89	410	67	56
9	47 691.62	431	78.50	30
10	43 827.01	910	67.59	9
均值	44 200.45	599.8	69.59	31

从表 4 和表 5 可以看出，基于条件随机场的"食品名称"和"具体因素"识别模型性能要优于基于最大熵模型的性能。条件随机场模型的 F 值最低为 90.06%，最高达到 91.94%，平均为 90.88%；最大熵模型的 F 值最高仅为 73.38%，平均仅为 70.48%。

从表 6 可以看出，在训练和测试的耗时来看，最大熵模型要优于条件随机场模型。前者一次训练与测试耗时在 100 秒左右，而后者需要约 50 000 秒左右。

由于后续研究更注重"食品名称"和"具体因素"识别的性能而非训练耗时的长短，因此本文选择条件随机场模型进行"食品名称"和"具体因素"的识别。对条件随机场模型所识别出来的"食品名称"和"具体因素"进行简单分析，发现识别错误较多的"食品名称"和"具体因素"主要是长度过程，比如"食品名称"和"具体因素"，比如"副溶血弧菌细菌"、"乔家栅高庄馒头"、"兽用加硒腐殖酸钠"、"受蜡样芽孢杆菌污染"、"汪氏蜂胶软胶囊"这些实体中要么含有难以识别的多重地名和形容词，比如"乔家"、"栅高"，要么姓名与名词组合，比如"汪氏"、"蜂胶"。这些实体中的复杂构成成分影响了对食品安全事件语料中实体识别的准确率和召回率。

在知网平台上对 2005 年所报道的任意食品安全事件新闻进行自动抓取并完成对文本的清洗。本文开发相应的软件，调用已经构建的食品安全事件实体抽取模型，完成对新闻报道中有关食品安全名称和具体因素的实体抽取。知网数据爬取功能截图如图 3 所示。



图 3　知网数据爬取功能截图

实体抽取功能截图如图 4 所示。

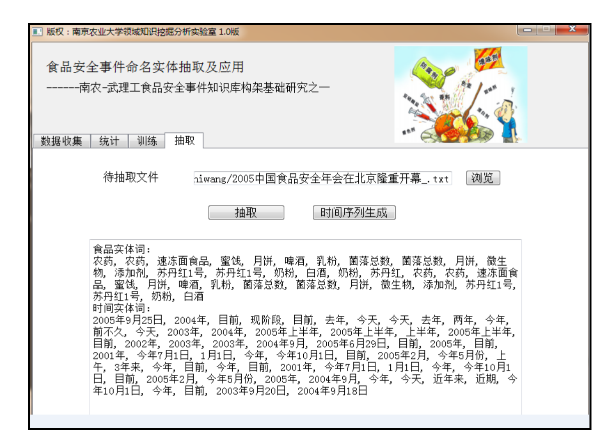


图 4　实体抽取功能截图

5　结　语

本文所界定的"食品名称"和"具体因素"实体的自动标注对于构建食品安全事件知识库和挖掘食品安全应对策略起到了充当基础资源的作用。在已标注"食品名称"和"具体因素"实体的语料基础上，通过统计实体的内外特征，构建了实体抽取的机器学习模型。从开放测试的结果观察，本文所构建的实体抽取模型整体性能较为突出，基本达到了实用的目标。在后续的研究中，一方面要在 1995 年–2004 年的时间跨度的语料上使用该模型进行具体的应用推广，另一方面结合模型的整体性能表现，通过融合新的特征改进已有模型的性能。

参考文献：

[1] 掷出窗外 [EB/OL]. [2014-02-18]. http://www.zccw.info/. (Zhi Chu Chuang Wai [EB/OL]. [2014-02-18]. http://www.zccw.info/.)

[2] 张慕洁，沈建华. 关于处置食品药品安全突发事件中信息公开的思考[J]. 上海食品药品监管情报研究，2012(2): 45-49. (Zhang Mujie, Shen Jianhua. About the Disposal of the Food and Drug Safety Incident Information to the Public Thinking about the Disposal of the Food and Drug Safety Incident Information[J]. Shanghai Food and Drug Information Research, 2012(2): 45-49.)

[3] 马颖，张园园，宋文广. 食品行业事件风险感知的传染病模型研究[J]. 科研管理，2013, 34(9): 123-130. (Ma Ying, Zhang Yuanyuan, Song Wenguang. Research on Epidemic Model of Emergency Events Risk Perception in Food Industry [J]. Science Research Management, 2013, 34(9): 123-130.)

[4] 陈宇，郑德权，赵铁军. 基于 Deep Belief Nets 的中文名实

体关系抽取[J]. 软件学报, 2012, 23(10): 2572-2585. (Chen Yu, Zheng Dequan,Zhao Tiejun. Chinese Relation Extraction Based on Deep Belief Nets[J]. Journal of Software, 2012, 23(10): 2572-2585.)

[5] 邵发, 黄银阁, 周兰江, 等. 基于实体消歧的中文实体关系抽取[J]. 山东大学学报: 工学版, 2014, 44(6): 32-37. (Shao Fa, Huang Yin'ge, Zhou Lanjiang, et al. Chinese Entity Relation Extraction Based on Entity Disambiguation[J]. Journal of Shandong University: Engineering Science, 2014, 44(6): 32-37.)

[6] 许华, 刘茂福, 姜丽, 等. 基于语言规则的病症菌实体抽取[J]. 武汉大学学报: 理学版, 2015, 61(2): 51-55. (Xu Hua, Liu Maofu, Jiang Li, et al. Disease and Bacteria Entity Extraction Based on Linguistic Rule[J].Journal of Wuhan University: Natural Science Edition, 2015, 61(2): 51-55.)

[7] 魏秀卓. 食品投诉文本敏感词汇抽取研究[D]. 长春: 东北师范大学, 2015. (Wei Xiuzhuo. Food Complaint Text Sensitive Words Extraction Research [D]. Changchun: Northeast Normal University, 2015.)

[8] 高蕊. 基于本体的食品投诉文本危害信息抽取研究[D]. 长春: 东北师范大学, 2011. (Gao Rui. Ontology-based Hazard Information Extraction from Chinese Food Complaint Documents[D]. Changchun: Northeast Normal University, 2011.)

[9] 李丽双, 党延忠, 张婧, 等. 基于条件随机场的汽车领域术语抽取[J]. 大连理工大学学报, 2013, 53(2): 267-272. (Li Lishuang, Dang Yanzhong, Zhang Jing, et al. Automotive Term Extraction Based on Conditional Random Fields[J]. Journal of Dalian University of Technology, 2013, 53(2): 267-272.)

[10] 王文龙, 王东波. 面向项目申请书的命名实体抽取模型构建研究[J]. 情报资料工作, 2015(1): 30-34. (Wang Longwen, Wang Dongbo. Project Application-oriented Named Entity Extraction Model Construction [J]. Information and Documentation Services, 2015(1): 30-34.)

[11] 刘凯, 周雪忠, 于剑, 等. 基于条件随机场的中医临床病历命名实体抽取[J]. 计算机工程, 2014, 40(9): 312-316. (Liu Kai, Zhou Xuezhong, Yu Jian, et al. Named Entity Extraction of Traditional Chinese Medicine Medical Records Based on Conditional Random Field[J]. Computer Engineering, 2014, 40(9): 312-316.)

[12] 吴云芳. 面向语言信息处理的现代汉语并列结构研究[M]. 北京: 北京师范大学出版社, 2004. (Wu Yunfang. Researches of Modern Chinese Coordinate Construction for Language Information Processing[M]. Beijing: Beijing Normal University Press, 2004.)

[13] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// Proceedings of the 18th International Conference on Machine Learning. 2001: 282-289.

[14] McCallum A, Freitag D, Pereira F. Maximum Entropy Markov Models for Information Extraction and Segmentation [C]//Proceedings of the 17th International Conference on Machine Learning. 2000: 591-598.

作者贡献声明:

王东波: 论文的框架搭建, 论文撰写、修订;

吴毅: 模型的训练和论文撰写;

叶文豪: 数据的标注;

刘睿伦: 语料的预处理。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: db.wang@njau.edu.cn。

[1] 王东波, 吴毅, 叶文豪, 刘睿伦. Event statistics programming. 基于食品安全事件语料库的实体统计程序.

[2] 王东波, 吴毅, 叶文豪, 刘睿伦. Event extracting platform. 基于条件随机场模型的实体抽取平台.

收稿日期: 2016-08-03
收修改稿日期: 2016-12-07

Extracting Events of Food Safety Emergencies with Characteristics Knowledge

Wang Dongbo[1,2]　Wu Yi[1]　Ye Wenhao[1]　Liu Ruilun[1]
[1](College of Information and Technology, Nanjing Agricultural University, Nanjing 210095, China)
[2](Research Center for Correlation of Domain Knowledge, Nanjing Agricultural University, Nanjing 210095, China)

Abstract: [**Objective**] This paper aims to extract the events of food safety emergencies from large food safety emergencies. [**Methods**] First, we built the food safety emergency corpus based on the past events, as well as the data acquisition, labeling, and organization methods of information science. Then, we extracted the corresponding events with the help of conditional random field model, and the distribution characteristics knowledge of the food safety emergencies. [**Limitations**] We might not be able to apply the feature template created by this research to other fields. [**Results**] We examined the proposed model with a food safety emergency corpus of 15 million Chinese words, and the F value of this model reached 91.94%. [**Conclusions**] It is feasible for us to extract the events from food safety emergency corpus with the help of conditional random field model.

Keywords: Characteristics Knowledge　Conditional Random Fields　Event　Food Safety Emergency

OCLC 发布研究报告探讨研究数据管理的现实状况

2017 年 3 月, OCLC 发布一项新的研究报告, 题为“研究数据管理(Research Data Management, RDM)服务空间之旅”, 概述了 RDM 服务空间的情况, 为世界 4 所大学进一步探索 RDM 奠定了基础。

该报告是“研究数据管理现实”系列报告中的第一部分, 这一系列报告重点分析了 4 个机构的决策情况, 这 4 个机构在面对研究型大学 RDM 服务规划、开发和部署时做出不同的选择。

该报告首先对爱丁堡大学(英国)、伊利诺伊大学香槟分校(美国)、蒙纳士大学(澳大利亚)和瓦赫宁根大学(荷兰)等 4 所大学进行案例分析, 研究这些机构的 RDM 能力。

报告撰写人解释说: “研究数据管理已经成为高等教育中十分重要的一个领域, 需要对服务、资源和基础设施进行大量投资, 以支持研究人员的数据管理需求。该报告是 OCLC Research 的一系列报告中的第一篇, 研究了高等教育机构在构建或获取 RDM 能力方面所面临的背景、影响和选择, 也即, 在支持新兴数据管理实践所需的基础设施、服务和其他资源时所面临的背景、影响和选择。”

除 4 项深入案例研究外, 报告还在各种国家环境中, 对北美、欧洲和澳大利亚的十几个研究型大学进行了 RDM 服务审查, 发现 RDM 服务可以分为三类:

(1) 教育类: 旨在教育研究人员和其他利益攸关方负责任地管理其数据以及安排长期保存的重要性, 甚至是必要性;

(2) 专业类: 这些服务为遇到具体研究数据管理问题的研究人员提供决策支持和定制解决方案;

(3) 保存类: 提供支持整个研究周期的数据管理的相关技术基础设施和相关服务。

“研究数据管理服务空间之旅”探索了这三个类别, 为整个系列报告提供了一个框架, 并对该系列中的下一个报告进行了预告。从 OCLC 研究网站可下载该报告。

(编译自: http://www.oclc.org/en/news/releases/2017/201708dublin.html)

(本刊讯)